Homework 6: Nice Demo City, But Will It Scale?

36-402, Advanced Data Analysis

Due at the start of class, 21 February 2011

For data-collection purposes, urban areas of the United States are divided into several hundred "Metropolitan Statistical Areas" based on patterns of residence and commuting; these cut across the boundaries of legal cities and even states. In the last decade, the U.S. Bureau of Economic Analysis has begun to estimate "gross metropolitan products" for these areas — the equivalent of gross national product, but for each metropolitan area. (See Homework 2 for the definition of "gross national product".) Even more recently, it has been claimed that these gross metropolitan products show a simple quantitative regularity, called "supra-linear power-law scaling". If Y is the gross metropolitan product in dollars, and N is the number of people in the city, then, the claim goes,

$$Y \approx cN^b \tag{1}$$

where the exponent $b > 1$ and the scale factor $c > 0$. This homework will use the tools built so far to test this hypothesis.

1. (15 points) A metropolitan area's gross per capita product is $y = Y/N$. Show that if Eq. 1 holds, then

   $$\log y \approx \beta_0 + \beta_1 \log N$$

   How are β_0 and β_1 related to c and b?

2. (15 points) The data files `gmp_2006.csv` and `pcgmp_2006.csv` on the class website contain the total gross metropolitan product (Y) in millions of dollars, and the per capita gross metropolitan product (y) in dollars, for all metropolitan areas in the US in 2006. Read them in and use them to calculate the metropolitan populations (N). If it's done correctly, then running `summary` on the population figures should give

   ```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
   54980   135600   231500   680900   530900 18850000
   ```

   (Your exact results may differ very slightly because of rounding and display settings.) What is the variance of $\log y$?

3. (20 points) *Estimating the power-law scaling model.* Use `lm` to linearly regress log per capita product, $\log y$, on log population, $\log N$. How does estimating this statistical model relate to Equation 1? What are the estimated coefficients? Are they compatible with the idea of supra-linear scaling? What is the mean squared error for $\log y$?

4. (15 points) Plot per capita product y against N, along with the fitted power-law relationship from problem 3. (Be careful about logs!)

5. (15 points) Fit a non-parametric smoother to $\log y$ and $\log N$. (You can use kernel regression, a spline, or any other non-parametric smoother.) What is the mean squared error for $\log y$? Describe, in words, how this curve compares to the power-law model from problem 3.

6. (20 points) Using the method from lecture 10, section 1, test whether the power-law relationship is correctly specified. What is the p-value? What do you conclude about the validity of the power-law model, based not just on this problem but the previous ones as well?